CPSC 540: Machine Learning

Density Estimation

Mark Schmidt

University of British Columbia

Winter 2018

Last Time: Density Estimation

- The next topic we'll focus on is density estimation:

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \quad \tilde{X} = \begin{bmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

  - What is probability of x^i for a generic feature vector x^i?

- For the training data this is easy:
  - Set $p(x^i)$ to "number of times x^i is in the training data" divided by n.

- We're interested in the probability of test data,
  - What is probability of seeing feature vector $\tilde{x}^i$ for a new example i.

Bernoulli Distribution on Binary Variables

- Let's start with the simplest case: $x^i \in \{0, 1\}$ (e.g., coin flips),

$$X = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

- For IID data the only choice is the Bernoulli distribution:

$$p(x^i = 1 \mid \theta) = \theta, \quad p(x^i = 0 \mid \theta) = 1 - \theta.$$

- We can write both cases

$$p(x^i|\theta) = \theta^{\mathcal{I}[x^i=1]}(1-\theta)^{\mathcal{I}[x^i=0]}, \text{ where } \mathcal{I}[y] = \begin{cases} 1 & \text{if } y \text{ is true} \\ 0 & \text{if } y \text{ is false} \end{cases}.$$

Maximum Likelihood with Bernoulli Distribution

- MLE for Bernoulli likelihood is

$$
\begin{aligned}
\underset{0\leq\theta\leq 1}{\operatorname{argmax}}\, p(X|\theta) &= \underset{0\leq\theta\leq 1}{\operatorname{argmax}} \prod_{i=1}^{n} p(x^i|\theta) \\
&= \underset{0\leq\theta\leq 1}{\operatorname{argmax}} \prod_{i=1}^{n} \theta^{\mathcal{I}[x^i=1]}(1-\theta)^{\mathcal{I}[x^i=0]} \\
&= \underset{0\leq\theta\leq 1}{\operatorname{argmax}} \underbrace{\theta^1\theta^1\cdots\theta^1}_{\text{number of } x_i = 1} \underbrace{(1-\theta)(1-\theta)\cdots(1-\theta)}_{\text{number of } x_i = 0} \\
&= \underset{0\leq\theta\leq 1}{\operatorname{argmax}}\, \theta^{N_1}(1-\theta)^{N_0},
\end{aligned}
$$

  where N_1 is count of number of 1 values and N_0 is the number of 0 values.
- If you equate the derivative of the log-likelihood with zero, you get $\theta = \frac{N_1}{N_1+N_0}$.
- So if you toss a coin 50 times and it lands heads 24 times, your MLE is $24/50$.

Multinomial Distribution on Categorical Variables

- Consider the multi-category case: $x^i \in \{1, 2, 3, \ldots, k\}$ (e.g., rolling di),

$$X = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 3 \\ 1 \\ 2 \end{bmatrix}.$$

- The categorical distribution is

$$p(x^i = c | \theta_1, \theta_2, \ldots, \theta_k) = \theta_c,$$

  where $\sum_{c=1}^{k} \theta_c = 1$.
- We can write this for a generic x as

$$p(x^i | \theta_1, \theta_2, \ldots, \theta_k) = \prod_{c=1}^{k} \theta_c^{\mathcal{I}[x^i = c]}.$$

Multinomial Distribution on Categorical Variables

- Using Lagrange multipliers (bonus) to handle constraints, the MLE is

$$\theta_c = \frac{N_c}{\sum_{c'} N_{c'}}. \quad (\text{“fraction of times you rolled a 4”})$$

- If we never see category 4 in the data, should we assume $\theta_4 = 0$?
    - If we assume $\theta_4 = 0$ and we have a 4 in test set, our test set likelihood is 0.
- To leave room for this possibility we often use “Laplace smoothing”,

$$\theta_c = \frac{N_c + 1}{\sum_{c'}(N_{c'} + 1)}.$$

- This is like adding a “fake” example to the training set for each class.

MAP Estimation with Bernoulli Distributions

- In the binary case, a generalization of Laplace smoothing is

$$\theta = \frac{N_1 + \alpha - 1}{(N_1 + \alpha - 1) + (N_0 + \beta - 1)},$$

- We get the MLE when $\alpha = \beta = 1$, and Laplace smoothing with $\alpha = \beta = 2$.
- This is a MAP estimate under a beta prior,

$$p(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where the beta function B makes the probability integrate to one.

$$\text{We want } \int_\theta p(\theta|\alpha,\beta)d\theta = 1, \quad \text{so define} \quad B(\alpha,\beta) = \int_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta.$$

- Note that $B(\alpha,\beta)$ is constant in terms of θ, it doesn't affect MAP estimate.

MAP Estimation with Categorical Distributions

- In the categorical case, a generalization of Laplace smoothing is

$$\theta_c = \frac{N_c + \alpha_c - 1}{\sum_{c'=1}^{k}(N_{c'} + \alpha_{c'} - 1)},$$

which is a MAP estimate under a Dirichlet prior,

$$p(\theta_1, \theta_2, \ldots, \theta_k | \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{c=1}^{k} \theta_c^{\alpha_c - 1},$$

where $B(\alpha)$ makes the multivariate distribution integrate to 1 over θ,

$$B(\alpha) = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_{k-1}}\int_{\theta_k}\prod_{c=1}^{k}\left[\theta_c^{\alpha_c-1}\right] d\theta_k d\theta_{k-1}\cdots d\theta_2 d\theta_1.$$

- Because of MAP-regularization connection, Laplace smoothing is regularization.

General Discrete Distribution

- Now consider the case where $x^i \in \{0,1\}^d$ (e..g, words in e-mails):

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

- Now there are 2^d possible values of x^i.
  - Can't afford to even store a θ for each possible x^i.
  - With n training examples we see at most n unique x^i values.
  - But unless we have a small number of repeated x^i values, we'll hopelessly overfit.
- With finite dataset, we'll need to make assumptions...

Product of Independent Distributions

- A common assumption is that the variables are independent:

$$p(x_1^i, x_2^i, \ldots, x_d^i|\Theta) = \prod_{j=1}^{d} p(x_j^i \mid \theta_j).$$

- Now we just need to model each column of X as its own dataset:

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \quad \rightarrow \quad X_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad X_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \quad X_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

- A big assumption, but now you can fit Bernoulli for each variable.
  - We did this in CPSC 340 for naive Bayes.

Density Estimation and Fundamental Trade-off

- "Product of independent" distributions:
  - Easily estimate each θ_c but can't model many distributions.
- General discrete distribution:
  - Hard to estimate 2^d parameters but can model any distribution.

- An unsupervised version of the fundamental trade-off:
  - Simple models often don't fit the data well but don't overfit much.
  - Complex models fit the data well but often overfit.

- We'll consider models that lie between these extremes:
  1. Mixture models.
  2. Graphical models.
  3. Boltzmann machines.

Outline

Univariate Gaussian

- Consider the case of a continuous variable $x \in \mathbb{R}$:

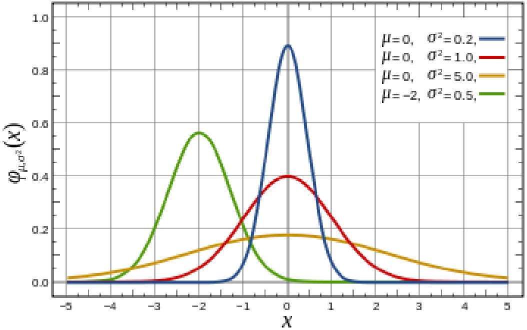
$$X = \begin{bmatrix} 0.53 \\ 1.83 \\ -2.26 \\ 0.86 \end{bmatrix}.$$

- Even with 1 variable there are many possible distributions.
- Most common is the Gaussian (or "normal") distribution:

$$p(x^i|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x^i-\mu)^2}{2\sigma^2}\right) \quad \text{or} \quad x^i \sim \mathcal{N}(\mu,\sigma^2),$$

for $\mu \in R$ and $\sigma > 0$.

Univariate Gaussian



https://en.wikipedia.org/wiki/Gaussian_function

- Mean parameter μ controls location of center of density.
- Variance parameter σ^2 controls how spread out density is.

Univariate Gaussian

- Why use the Gaussian distribution?
  - Data might actually follow Gaussian.
    - Good justification if true, but usually false.
  - Central limit theorem: mean estimators converge in distribution to a Gaussian.
    - Bad justification: doesn't imply data distribution converges to Gaussian.
  - Distribution with maximum entropy that fits mean and variance of data (bonus).
    - "Makes the least assumptions" while matching first two moments of data.
    - But for complicated problems, just matching mean and variance isn't enough.
  - Closed-form maximum likelihood esitmate (MLE).
    - MLE for the mean is the mean of the data ("sample mean" or "empirical mean").
    - MLE for the variance is the variance of the data ("sample variance").
    - "Fast and simple".

Univariate Gaussian

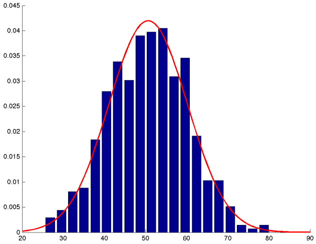
- Gaussian likelihood for an example x^i is
$$p(x^i|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x^i-\mu)^2}{2\sigma^2}\right).$$
- So the negative log-likelihood for n IID examples is
$$-\log p(X|\mu,\sigma^2) = -\sum_{i=1}^n \log p(x^i|\mu,\sigma^2) = \frac{1}{2\sigma^2}\sum_{i=1}^n (x^i-\mu)^2 + n\log(\sigma) + \text{const}.$$
- Setting derivative with respect to μ to 0 gives MLE of
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x^i. \quad \text{(for any } \sigma > 0\text{)}$$
- Plugging in $\hat{\mu}$ and setting derivative with respect to σ to zero gives
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x^i-\hat{\mu})^2. \quad \text{(if this zero, the NLL is unbounded and MLE doesn't exist).}$$

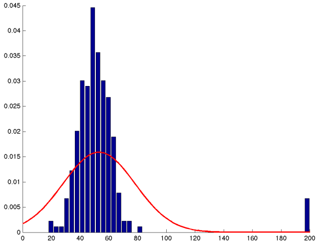
Alternatives to Univariate Gaussian

- Why not the Gaussian distribution?
  - Negative log-likelihood is a quadratic function of μ,

$$-\log p(X|\mu,\sigma^2) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^i-\mu)^2 + n\log(\sigma) + \text{const.}$$

so as with least squares the Gaussian is not robust to outliers.

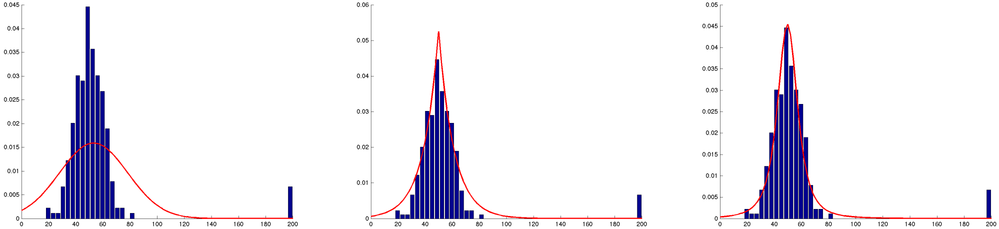




- "Light-tailed": assumes most data is really close to mean.

Alternatives to Univariate Gaussian

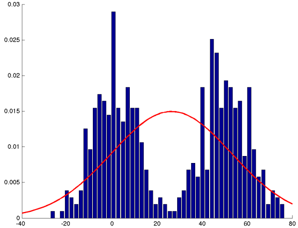
- Robust: Laplace distribution or student's t-distribution



- "Heavy-tailed": has non-trivial probability that data is far from mean.

Alternatives to Univariate Gaussian

- Gaussian distribution is unimodal.



- Laplace and student t are also unimodal so don't fix this issue.
    - Next time we'll discuss "mixture models" that address this.

Product of Independent Gaussians

- If we have d variables, we could make each follow an independent Gaussian,

$$x_j^i \sim \mathcal{N}(\mu_j, \sigma_j^2),$$

- In this case the joint density over all d variables is

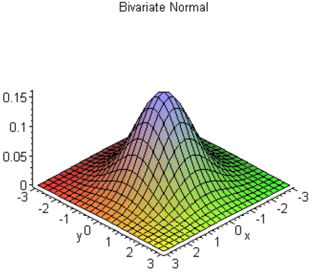
$$\begin{aligned}
\prod_{j=1}^{d} p(x_j^i|\mu_j,\sigma_j^2) &\propto \prod_{j=1}^{d} \exp\left(-\frac{(x_j^i-\mu_j)^2}{2\sigma_j^2}\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{d}\frac{1}{\sigma_j^2}(x_j^i-\mu_j)^2\right) && (e^a e^b = e^{a+b}) \\
&= \exp\left(-\frac{1}{2}(x^i-\mu)^T\Sigma^{-1}(x-\mu)\right) && \left(\sum_{j=1}^{d} v_j w_j^2 = w^T V^{\frac{1}{2}} V^{\frac{1}{2}} w\right).
\end{aligned}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_d)$ and Σ is diagonal with diagonal elements σ_j^2.

- This is a special case of a multivariate Gaussian with a diagonal covariance Σ.

Multivariate Gaussian Distribution

- The generalization to multiple variables is the multivariate normal/Gaussian,



http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html

- The probability density is given by

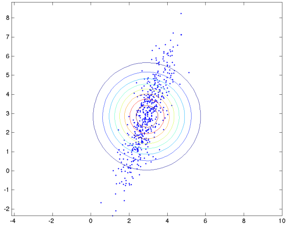
$$p(x^i|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^T \Sigma^{-1}(x^i - \mu)\right), \quad \text{or } x^i \sim \mathcal{N}(\mu, \Sigma),$$

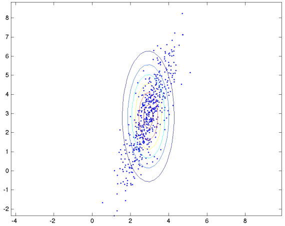
where $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma \succ 0$, and $|\Sigma|$ is the determinant.

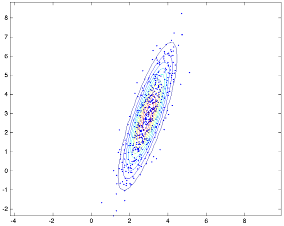
Product of Independent Gaussians

- The effect of a diagonal Σ on the multivariate Gaussian:
  - If $\Sigma = \alpha I$ the level curves are circles: 1 parameter.
  - If $\Sigma = D$ (diagonal) then axis-aligned ellipses: d parameters.
  - If Σ is dense they do not need to be axis-aligned: $d(d+1)/2$ parameters.

(by symmetry, we only need upper-triangular part of Σ)







- Diagonal Σ assumes features are independent, dense Σ models dependencies.

Maximum Likelihood Estimation in Multivariate Gaussians

- With a multivariate Gaussian we have

$$p(x^i|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right),$$

  so up to a constant our negative log-likelihood is

$$\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu) + \frac{n}{2}\log|\Sigma|.$$

- This is a strongly-convex quadratic in μ, so setting gradient to zero

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x^i,$$

  which is the unique solution (strong-convexity is due to $\Sigma \succ 0$).
  - MLE for μ is the average along each dimension, and it doesn't depend on Σ.

Summary

- Density estimation: unsupervised modelling of probability of feature vectors.
- Categorical distribution for modeling discrete data.
- Product of independent distributions is simple/crude density estimation method.
- Gaussian distribution is a common distribution with many nice properties.
  - Closed-form MLE.
  - But unimodal and not robust.
- Next time: going beyond Gaussians.

Lagrangian Function for Optimization with Equality Constraints

- Consider minimizing a differentiable f with linear equality constraints,

$$\underset{Aw=b}{\text{argmin}}\, f(w).$$

- The Lagrangian of this problem is defined by

$$L(w,v) = f(w) + v^T(Aw - b),$$

  for a vector $v \in \mathbb{R}^m$ (with A being m by d).
- At a solution of the problem we must have

$$\nabla_w L(w,v) = \nabla f(w) + A^T v = 0 \qquad \text{(gradient is orthogonal to constraints)}$$
$$\nabla_v L(w,v) = Aw - b = 0 \qquad \text{(constraints are satisfied)}$$

- So solution is stationary point of Lagrangian.

Lagrangian Function for Optimization with Equality Constraints

- Scans from Bertsekas discussing Lagrange multipliers (also see CPSC 406).

3.1 NECESSARY CONDITIONS FOR EQUALITY CONSTRAINTS

In this section we consider problems with equality constraints of the form

$$\begin{aligned}&\text{minimize } f(x)\\&\text{subject to } h_i(x)=0, \quad i=1,\ldots,m.\end{aligned}\tag{ECP}$$

We assume that $f: \Re^n \mapsto \Re$, $h_i: \Re^n \mapsto \Re$, $i=1,\ldots,m$, are continuously differentiable functions. *All the necessary and the sufficient conditions of this chapter relating to a local minimum can also be shown to hold if f and h_i are defined and are continuously differentiable within just an open set containing the local minimum.* The proofs are essentially identical to those given here.

For notational convenience, we introduce the constraint function $h: \Re^n \mapsto \Re^m$, where

$$h=(h_1,\ldots,h_m).$$

We can then write the constraints in the more compact form

$$h(x)=0. \tag{3.1}$$

Our basic Lagrange multiplier theorem states that for a given local minimum x^*, there exist scalars $\lambda_1,\ldots,\lambda_m$, called *Lagrange multipliers*, such that

$$\nabla f(x^*)+\sum_{i=1}^m \lambda_i \nabla h_i(x^*)=0. \tag{3.2}$$

There are two ways to interpret this equation:

(a) The cost gradient $\nabla f(x^*)$ belongs to the subspace spanned by the constraint gradients at x^*. The example of Fig. 3.1.1 illustrates this interpretation.

(b) The cost gradient $\nabla f(x^*)$ is orthogonal to the subspace of *first order feasible variations*

$$V(x^*)=\{\Delta x \mid \nabla h_i(x^*)'\Delta x=0,\ i=1,\ldots,m\}.$$

This is the subspace of variations Δx for which the vector $x=x^*+\Delta x$ satisfies the constraint $h(x)=0$ up to first order. Thus, according to the Lagrange multiplier condition of Eq. (3.2), at the local minimum x^*, the first order cost variation $\nabla f(x^*)'\Delta x$ is zero for all variations

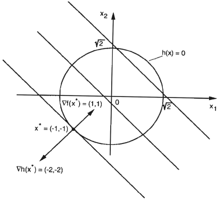


Figure 3.1.1. Illustration of the Lagrange multiplier condition (3.1) for the problem

$$\begin{aligned}&\text{minimize } x_1+x_2\\&\text{subject to } x_1^2+x_2^2=2.\end{aligned}$$

At the local minimum $x^*=(-1,-1)$, the cost gradient $\nabla f(x^*)$ is normal to the constraint surface and is therefore, collinear with the constraint gradient $\nabla h(x^*)=(-2,-2)$. The Lagrange multiplier is $\lambda=1/2$.

Δx in this subspace. This statement is analogous to the "zero gradient condition" $\nabla f(x^*)=0$ of unconstrained optimization.

Here is a formal statement of the main Lagrange multiplier theorem:

Proposition 3.1.1: (Lagrange Multiplier Theorem – Necessary Conditions) Let x^* be a local minimum of f subject to $h(x)=0$, and assume that the constraint gradients $\nabla h_1(x^*),\ldots,\nabla h_m(x^*)$ are linearly independent. Then there exists a unique vector $\lambda^*=(\lambda_1^*,\ldots,\lambda_m^*)$, called a *Lagrange multiplier vector*, such that

$$\nabla f(x^*)+\sum_{i=1}^m \lambda_i^* \nabla h_i(x^*)=0. \tag{3.3}$$

If in addition f and h are twice continuously differentiable, we have

Lagrangian Function for Optimization with Equality Constraints

- We can use these optimality conditions,

$$\nabla_w L(w,v) = \nabla f(w) + A^T v = 0 \qquad \text{(gradient is orthogonal to constraints)}$$
$$\nabla_v L(w,v) = Aw - b = 0 \qquad \text{(constraints are satisfied)}$$

  to solve some constrained optimization problems.

- A typical approach might be:
  1. Solve for w in the equation $\nabla_w L(w,v) = 0$ to get $w = g(v)$ for some function g.
  2. Plug this $w = g(v)$ into the the equation $\nabla_v L(w,v) = 0$ to solve for v.
  3. Use this v in $g(v)$ to get the optimal w.

- But note that these are necessary conditions (may need to check it's a min).

Maximum Entropy and Gaussian

- Consider trying to find the PDF $p(x)$ that
  1. Agrees with the sample mean and sample covariance of the data.
  2. Maximizes entropy subject to these constraints,

$$\max_p \left\{ -\int_{-\infty}^{\infty} p(x) \log p(x) dx \right\}, \quad \text{subject to } \mathbb{E}[x] = \mu,\ \mathbb{E}[(x-\mu)^2] = \sigma^2.$$

- Solution is the Gaussian with mean μ and variance σ^2.
  - Beyond fitting mean/variance, Gaussian makes fewest assumptions about the data.
- This is proved using the convex conjugate (see duality lecture).
  - Convex conjugate of Gaussian negative log-likelihood is entropy.
  - Same result holds in higher dimensions for multivariate Gaussian.